Title

PhosBoost: Improved phosphorylation prediction recall using gradient boosting and protein language models.

Journal

Plant Direct, 7(12)

Authors

Poretsky, Elly

Andorf, Carson

Sen, Taner

Publication Date

2023-12-01

DOI

10.1002/pld3.554

Peer reviewed

Received: 10 August 2023 | Revised: 20 November 2023 | Accepted: 26 November 2023

DOI: 10.1002/pld3.554

RESEARCH ARTICLE

PhosBoost: Improved phosphorylation prediction recall using gradient boosting and protein language models

Elly Poretsky[1] | Carson M. Andorf[2,3] | Taner Z. Sen[1,4]

[1]Agricultural Research Service, Crop Improvement and Genetics Research Unit, U.S. Department of Agriculture, Albany, CA, United States

[2]Agricultural Research Service, Corn Insects and Crop Genetics Research, U.S. Department of Agriculture, Ames, IA, United States

[3]Department of Computer Science, Iowa State University, Ames, IA, United States

[4]Department of Bioengineering, University of California, Berkeley, CA, United States

Correspondence
Taner Z. Sen, 800 Buchanan St., Albany, CA 94710.
Email: taner.sen@usda.gov

Funding information
This research was supported by the U.S. Department of Agriculture, Agricultural Research Service, Project Numbers 2030–21000-056-00D and 5030–21000-072-000D through the Crop Improvement and Genetics Research and Corn Insects and Crop Genetics Research Units. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and Employer.

Abstract

Protein phosphorylation is a dynamic and reversible post-translational modification that regulates a variety of essential biological processes. The regulatory role of phosphorylation in cellular signaling pathways, protein–protein interactions, and enzymatic activities has motivated extensive research efforts to understand its functional implications. Experimental protein phosphorylation data in plants remains limited to a few species, necessitating a scalable and accurate prediction method. Here, we present PhosBoost, a machine-learning approach that leverages protein language models and gradient-boosting trees to predict protein phosphorylation from experimentally derived data. Trained on data obtained from a comprehensive plant phosphorylation database, qPTMplants, we compared the performance of PhosBoost to existing protein phosphorylation prediction methods, PhosphoLingo and DeepPhos. For serine and threonine prediction, PhosBoost achieved higher recall than PhosphoLingo and DeepPhos (.78, .56, and .14, respectively) while maintaining a competitive area under the precision-recall curve (.54, .56, and .42, respectively). PhosphoLingo and DeepPhos failed to predict any tyrosine phosphorylation sites, while PhosBoost achieved a recall score of .6. Despite the precision-recall tradeoff, PhosBoost offers improved performance when recall is prioritized while consistently providing more confident probability scores. A sequence-based pairwise alignment step improved prediction results for all classifiers by effectively increasing the number of inferred positive phosphosites. We provide evidence to show that PhosBoost models are transferable across species and scalable for genome-wide protein phosphorylation predictions. PhosBoost is freely and publicly available on GitHub.

1 | INTRODUCTION

Protein phosphorylation is one of the most widespread and important post-translational modifications that play a pivotal role in the regulation of protein function and cellular pathways (Nishi et al., 2014), and has been extensively studied in plants (Zhang et al., 2023). Through the covalent addition of a phosphate group to specific amino acids, predominantly serine (Ser, S), threonine (Thr, T), or tyrosine (Tyr, Y), protein phosphorylation alters various aspects of protein function, including activity, subcellular localization, stability, and interactions with other proteins or ligands (Álvarez-Salamero et al., 2017). Its involvement spans a wide array of cellular functions, such as cell signaling, metabolism, development, and resistance to biotic and abiotic stress (Chaudhuri et al., 2015; Dressano et al., 2020; Humphrey et al., 2015; Kim et al., 2009; Nishi et al., 2014; Oh et al., 2009; Ryu et al., 2007; Wang et al., 2013; Zhao & Guo, 2011). The dysregulation

of protein phosphorylation has been closely linked to various diseases and is also the target of effectors that can enhance pathogenic virulence (Ardito et al., 2017; Toruño et al., 2016). Often, multiple residues within the same protein will undergo phosphorylation that can have independent, synergistic, or antagonistic functions, revealing the existence of complex "phosphocodes" (Pejaver et al., 2014; Zhang et al., 2023) while crosstalk between different post-translational modifications can fine-tune cellular responses (Vu et al., 2018). Cross-species and -protein family comparisons have shown that functional phosphosites are more likely to be conserved, suggesting that phosphorylation information is transferable to the studies of other species (Chaudhuri et al., 2015). Thus, protein phosphorylation has emerged as a powerful target for developing disease treatments and manipulation for enhanced crop yields (Gough & Sadanandom, 2021; Yang et al., 2010).

Protein phosphorylation is among the most studied post-translational modifications using high-throughput mass-spectrometry-based methods, making a large amount of experimental protein phosphorylation data available from a variety of species, tissues, developmental stages, and conditions (Dou et al., 2021; Xue et al., 2022; Yu et al., 2023), but is relatively understudied in plant species besides *Arabidopsis thaliana* (Meng et al., 2022; Xue et al., 2022). Conservation of phosphosites within and across species has been observed offering a complementary approach for improving phosphosite detection and annotation based on straightforward sequence similarity approaches (Amanchy et al., 2007; Chaudhuri et al., 2015; Maathuis, 2008; Tan et al., 2009). The large increase in the amount of experimental protein phosphorylation data facilitated the development of a variety of machine- and deep-learning-based protein phosphorylation classifications (Luo et al., 2019; Wang et al., 2017; Zuallaert et al., 2022). For example, classical machine learning algorithms such as random forests (Ismail et al., 2016; Liu et al., 2022; Wei et al., 2017), support vector machines (Dou et al., 2014; Jamal et al., 2021), and gradient boosting trees (Maiti et al., 2020) have been used for protein phosphorylation prediction. More recently, deep learning algorithms such as convolutional neural networks (Guo et al., 2020; Luo et al., 2019; Wang et al., 2017; Zuallaert et al., 2022) and long short-term memory networks (Lv et al., 2021; Thapa et al., 2021) enabled learning directly from protein sequences, making substantial improvements in protein phosphorylation prediction.

In the absence of kinase specificity and targeted consensus sequences, kinase promiscuity adds to the challenge of protein phosphorylation prediction (Friso & Van Wijk, 2015). Different types of features have been used for protein phosphorylation prediction, often derived from biophysical properties, such as solvent accessibility and disorder score, and structural features, such as secondary structure (Y. Dou et al., 2014; Gao et al., 2010; Jamal et al., 2021). Improvements in deep learning classification methods enabled learning directly from the sequences surrounding the phosphosites without the need for complex feature representations (Luo et al., 2019; D. Wang et al., 2017; Wang et al., 2022). More recently, advances made in natural language processing enabled the development of protein language models (pLMs). By pre-training on vast numbers of protein sequences, pLMs learned inherent properties encoded within protein sequences, revolutionizing multiple fields of protein research (Bordin et al., 2023; Elnaggar et al., 2021; Ofer et al., 2021; Rives et al., 2021). Most importantly, pLMs capture both short and long-range functional and biophysical properties of each amino acid within a given protein as an encoded numerical vector, known as vector embeddings, that can be directly used by machine- and deep-learning classifiers (Littmann et al., 2021). The utilization of pLM-based vector embeddings has proven effective in predicting structural and biochemical properties such as secondary structures, solvent accessibility, and ligand binding residues (Ilzhöfer et al., 2022; Weissenow et al., 2022). The use of pLMs has also been used to develop new methods with improved performance for general and kinase-specific protein phosphorylation prediction (Zhou et al., 2023; Zuallaert et al., 2022). Recent advances in protein phosphorylation prediction, employing convolutional neural networks with pLM-based vector embeddings, demonstrated the potential of pLMs in predicting general protein phosphorylation, while also being applicable to other post-translational modifications (Zuallaert et al., 2022).

Compared to Ser and Thr phosphorylation, the prediction of Tyr phosphorylation poses additional challenges, partially due to a smaller amount of experimental data leading to a relatively higher label imbalance (Doll & Burlingame, 2015; La Fuente et al., 2009; Silva-Sanchez et al., 2015). Compared to animals, plants completely lack dedicated Tyr-specific kinases and rely on dual-specificity Ser/Thr and Tyr kinases for all Tyr phosphorylation, leading to a particularly imbalanced Tyr phosphorylation data in plants (Ghelis, 2011; La Fuente et al., 2009). Thus, while prediction of Tyr phosphorylation poses a challenge in both animals and plants, the challenge is expected to be more acute in plants. Despite the lack of dedicated Tyr kinases in plants, Tyr phosphorylation is known to regulate the function of multiple proteins involved in a variety of essential biological processes (Ghelis, 2011; Mühlenbeck et al., 2021). Consequently, improving the prediction accuracy of Tyr phosphorylation holds great potential for identifying functional phosphosites for a better understanding of cellular signaling and regulatory processes.

Here, we present the newly developed PhosBoost, a machine-learning classification method that uses pLMs with a stacking classifier composed of CatBoost (Prokhorenkova et al., 2018) gradient-boosting tree-based ensemble base classifiers to predict protein phosphorylation directly from protein sequences. We trained and evaluated PhosBoost on a large set of experimentally derived protein phosphorylation data obtained from the plant post-translational modification database, qPTMplants (Xue et al., 2022). Compared to existing protein phosphorylation prediction methods, PhosBoost consistently achieves higher recall, albeit at reduced or comparable precision for Ser/Thr prediction. PhosBoost provided both higher recall and precision for Tyr prediction. PhosBoost also produces probability scores that are more informative and better reflect the confidence in the protein phosphorylation prediction. To improve phosphosite annotation and reduce phosphosite label uncertainty, we supplemented PhosBoost predictions with a DIAMOND (Buchfink et al., 2021) pairwise

alignment analysis step that annotates phosphosites matching experimentally derived data. We provide evidence to show that PhosBoost models are transferable across species and scalable for genome-wide protein phosphorylation predictions, allowing for the incorporation of PhosBoost prediction results directly in the genome browser to facilitate accessibility. Our results show that PhosBoost is a competitive method for protein phosphorylation predictions, particularly when higher recall and genomic coverage of phosphosites are prioritized.

2 | RESULTS

2.1 | Overview of the protein phosphorylation data used to develop PhosBoost

Protein phosphorylation data was obtained from qPTMplants, a comprehensive database for high-throughput plant post-translational modification experimental data (Xue et al., 2022). The qPTMplants database includes protein phosphorylation data for over 30 plant species collected from different experiments, representing different organs, developmental stages, and conditions (Xue et al., 2022). Because of the predominance and relative saturation of the *A. thaliana* protein phosphorylation data (Figure S1A-C), we focused on the *A. thaliana* dataset for model training, hyperparameter tuning, and performance benchmarking. After collecting all experimentally derived positive phosphosites, the remaining Ser, Thr, and Tyr residues were collected to form the negative phosphosite dataset. Random stratification was used on the individual Ser, Thr, and Tyr datasets to generate the training, validation, and test sets using a 60%–20%–20% split, respectively. The Ser and Thr phosphosites were then combined for training and testing a combined binary S/T classification model.

2.2 | Description of the PhosBoost protein phosphorylation classifier

All the qPTMplants phosphoprotein sequences were used as an input for the ProtT5-XL-U50 pre-trained pLM (Elnaggar et al., 2021) that was selected based on the improved performance when used at protein classification tasks compared to other pre-trained pLMs (Zuallaert et al., 2022). From the encoded embedding vectors for each phosphoprotein, we extracted the Ser, Thr, and Tyr residue embedding vectors and the protein-wise average embedding vector (Figure 1a). PhosBoost was designed as a machine-learning stacking classifier that trains two separate binary classification models, one for Ser/Thr and one for Tyr prediction, using both the residue and protein embedding vector input data (Figure 1b). The stacking classifier consists of two CatBoost base classifiers, one classifier using balanced class weights based on label frequency and one CatBoost base classifier that uses equal class weights for the positive and negative labels (Figure 1b). The predicted probability results from the two stacked base classifiers were used as input features for the logistic regression metaclassifier that was trained using a 5-fold cross-validated prediction approach (Figure 1b).

2.3 | Assessing the performance of the stacking classifier

The performance of the PhosBoost stacking classifier was compared to the independently trained CatBoost classifiers, one trained with balanced class weights and one with equal class weights, trained and evaluated on the same *A. thaliana* qPTMplants dataset. Considering the S/T model, all three classifiers had a similar area under the receiver operating characteristic (AUROC) score of .88 (Figure 1c), while the balanced class weights CatBoost classifier had a lower area under precision-recall (AUPRC) score of .48 compared to .53 and .54 for the equal class weights CatBoost classifier and PhosBoost, respectively (Figure 1d). PhosBoost achieved a higher recall score at all probability thresholds (Figure 1e). A similar pattern was observed for the Y-model binary classifiers. While all models had similar AUROC scores of .70, .71, and .72 for PhosBoost, balanced class weights CatBoost classifier and equal class weights CatBoost classifier, respectively (Figure 1f), the balanced class weights CatBoost classifier had a lower AUPRC score of .48 compared to .53 and .54 for the equal class weights CatBoost classifier and PhosBoost, respectively (Figure 1g). PhosBoost achieved a higher recall score at all probability thresholds (Figure 1h). To provide additional support for using the PhosBoost stacking classifier approach, we conducted a similar analysis on an independent dataset, the Ramasamy22 protein phosphorylation dataset, obtained from the PhosphoLingo preprint (Zuallaert et al., 2022). As with the results obtained for the data trained on *A. thaliana* qPTMplants dataset, we observed that the PhosBoost stacking classifier provided an increased recall with no cost to precision for both the S/T and Y models (Figure S3).

2.4 | Evaluation of the predictive performance of PhosBoost in comparison to established protein phosphorylation prediction methods

We compared the performance of PhosBoost with two other existing protein phosphorylation classifiers, namely DeepPhos and PhosphoLingo (Luo et al., 2019; Zuallaert et al., 2022). The three methods were trained and evaluated on the same *A. thaliana* qPTMplants dataset. Considering the S/T model, PhosBoost performed better than DeepPhos but just under PhosphoLingo based on the AUROC scores (.86, .91, and, .89, respectively) (Figure 2a, Table 1) and AUPRC scores .54, .56, and .42, respectively, but lower than PhosphoLingo based on the F1 score, .43, .56, and .24, respectively (Figure 2b, Table 1). PhosBoost only achieved a higher precision score than PhosphoLingo when recall was below .41 (Figure 2b). PhosBoost achieved a higher recall score than PhosphoLingo and DeepPhos, .78, .56, and .14, respectively, and higher recall at all probability thresholds (Figure 2c, Table 1). Considering the Y model, we observed a similar pattern on

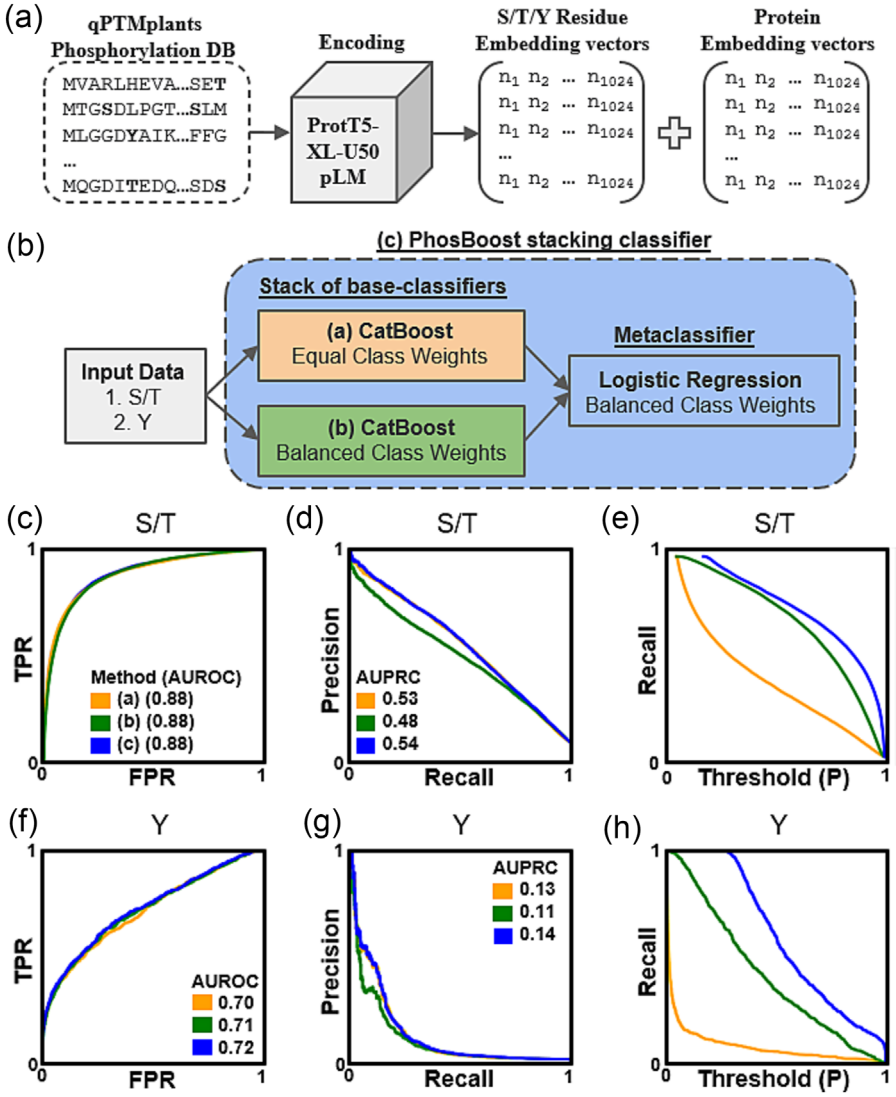

FIGURE 1 Overview of the PhosBoost protein phosphorylation classification workflow. (a) Protein phosphorylation data from the qPTMplants database were encoded by the pre-trained ProtT5-XL-U50 pLM to generate the embedding vector data for all serine (S), threonine (T), and tyrosine (Y) residues in addition to the protein-wise average embedding vector. (b) The input data for all *A. thaliana* phosphosites was used to generate two separate binary classifiers: (1) S/T model and (2) Y model, trained using a stacking classifier, termed PhosBoost. The stacking classifier is composed of two CatBoost base classifiers: (a) with equal class weights and (b) balanced class weights. A logistic regression metaclassifier combined the predicted probability scores of (a) and (b) to produce (c) PhosBoost. **(c-e)** The performances of the independently trained (a, orange) and (b, green) CatBoost classifiers were compared with the performance of the PhosBoost stacking classifier (c, blue), showing the receiver operating characteristic curve and area under receiver operating characteristic curve (AUROC) score, precision-recall curve and area under precision-recall curves (AUPRC) score, true positive rate (TPR), false positive rate (FPR), and probability (P) threshold, for the Ser/Thr model, and similarly **(F-H)** for the Tyr model.

the AUROC scores as for the S/T model, with PhosBoost having a higher score than DeepPhos but lower than PhosphoLingo, .72, .7, and .73, respectively (Figure 2d. Table 1). Unlike the S/T model, PhosBoost performed better than PhosphoLingo and DeepPhos based on the AUPRC scores, .14, .08, and .04, respectively, and based on the F1 score, .06, 0, and 0, respectively (Figure 2e, Table 1). Similar to the S/T model, PhosBoost achieved higher recall at all probability thresholds compared to DeepPhos and PhosphoLingo, and while the recall score for PhosBoost was .6, both DeepPhos and PhosphoLingo did not predict any Tyr phosphosites correctly (Figure 2f, Table 1).

Because neither PhosphoLingo nor DeepPhos correctly predicted Tyr phosphosites in the *A. thaliana* dataset, we decided to conduct an additional benchmark using a different dataset, namely the Ramasamy22 human phosphoproteomic dataset obtained from the PhosphoLingo preprint (Zuallaert et al., 2022). To compare the label imbalance between the two datasets, we analyzed the number of unique positive and negative phosphosites. For the negative phosphosites, we found that there were approximately twice as many Ser, Thr, and Tyr phosphosites in the *A. thaliana* qPTMplants dataset (Figure S2A). For the positive phosphosites, there were approximately twice as many Ser and Thr phosphosites in the *A. thaliana* qPTMplants dataset but approximately half the number of Tyr phosphosites (Figure S2B), explaining the observed similar label imbalance for the Ser and Thr phosphosites but higher Tyr label imbalance in the *A. thaliana* qPTMplants dataset (Figure S2C). Considering the S/T model, PhosBoost had a lower AUROC score than DeepPhos and PhosphoLingo, .90, .92, and .94, respectively (Figure S4A). Based on the AUPRC score, PhosBoost performed worse than PhosphoLingo but better than DeepPhos, .63, .71, and .51, respectively (Figure S4B). PhosBoost achieved a higher recall score of .83, compared to .64 and .26 for PhosphoLingo and DeepPhos, respectively, while achieving a higher recall score at all probability thresholds (Figure S4C). Considering the Y model, PhosBoost had a higher AUROC score than PhosphoLingo and DeepPhos, .93, .75, and .9, respectively (Figure S4D) and a higher AUPRC score, .55, .45, and .10, respectively (Figure S4E). PhosBoost achieved a higher recall score of .74, compared to .31 and .00 for PhosphoLingo and DeepPhos, respectively, while achieving a higher recall score at all probability thresholds (Figure S4F).

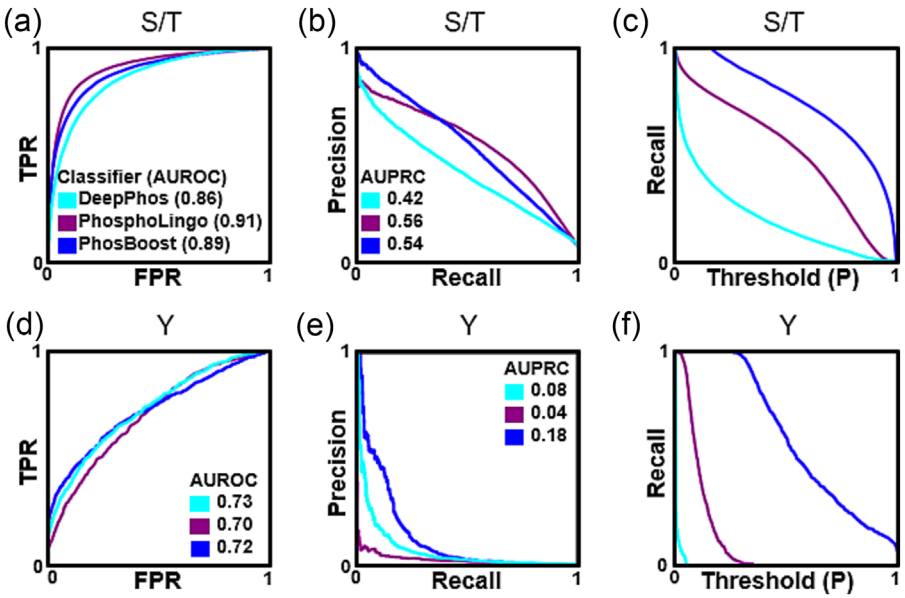

FIGURE 2 Comparing the predictive performance of PhosBoost, PhosphoLingo, and DeepPhos. (a-c) Comparison of the performance results for DeepPhos (teal), PhosphoLingo (purple), and PhosBoost (blue), showing the receiver operating characteristic curve and area under receiver operating characteristic curve (AUROC) score, precision-recall curve, and area under precision-recall curves (AUPRC) score, true positive rate (TPR), false positive rate (FPR), and probability (P) threshold, for the Ser/Thr model, and similarly (d-f) for the Tyr model.

TABLE 1 Comparing the classification metrics of PhosBoost with other existing phosphorylation prediction methods. The different protein phosphorylation methods were compared using different metrics, including the area under the receiver operating characteristic curve (AUROC), precision, recall, the area under the precision-recall curve (AUPRC), and F1 scores. The scores are presented for both the S/T and Y binary classifiers.

	AUROC		Precision		Recall		AUPRC		F1	
Method	**S/T**	**Y**	**S/T**	**Y**	**S/T**	**Y**	**S/T**	**Y**	**S/T**	**Y**
DeepPhos	.86	**.73**	**.67**	0	.14	0	.42	.08	.24	0
PhosphoLingo	**.91**	.70	.57	0	.56	0	**.56**	.04	**.56**	0
PhosBoost	.89	.72	.29	**0.03**	**.78**	**.60**	.54	**.14**	.43	**.06**

2.5 | Analysis of the predicted positive phosphosites suggests that PhosBoost provides more confident scores and outperforms PhosphoLingo and DeepPhos if recall is prioritized

While achieving higher performance at Tyr phosphosite prediction (Figure 2d-e, S4D-E, Table 1), PhosBoost performance at Ser/Thr prediction was better than DeepPhos but lower or comparable to PhosphoLingo (Figure 2b, S4B, Table 1). Despite the differences in performance, PhosBoost consistently obtained higher recall scores than PhosphoLingo and DeepPhos (Figure 2c, f, S4C, S4F, Table 1), pointing to a potential tradeoff between precision and recall. First, we plotted the confusion matrices to compare the results of the S/T models on the *A. thaliana* qPTMplants test set. The results show that the number of true positive (TP) Ser/Thr phosphosites predicted by PhosBoost, PhosphoLingo, and DeepPhos were 16,185, 11,580, and 2,967, respectively, and the number of false positives (FP) was 38,770, 8,795, and 1,492, respectively (Figure 3a-c). Focusing on the PhosBoost Y model results, due to the lack of correctly predicted Y phosphosites by DeepPhos and PhosphoLingo (Figure 2f, Table 1), the results show that 12,448 and 391 Tyr phosphosites were predicted as FP and TP, respectively (Figure S5A). Previously, we observed that PhosBoost had higher precision than PhosphoLingo at lower recall (Figure 2b), suggesting a difference in the distribution of the predicted probability scores. The split violin plots show the distribution of the TP and FP predicted phosphosites by PhosBoost, PhosphoLingo, and DeepPhos, for both the Ser and Thr phosphosites (Figure 3d). We observed that for all three classifiers, the predicted probability scores of the TPs were generally higher than of the FPs. The difference between the distributions was most distinct for PhosBoost (Figure 3d). Notably, the peak of the distribution for the TP predicted probability scores, for both the Ser and Thr results, was between .9 and 1 (Figure 3d). In contrast to the Ser and Thr results, we observed a bi-modal distribution for the true positive Tyr results, with one peak between .8 and 1 and another peak between .5 and .7 (Figure S5B). A similar pattern to the S/T results was observed for the S/T model trained on the Ramsamy22 dataset (Figure S6A-C) and the Y model (Figure 6d-f), although the distribution of predicted probability values for PhosBoost and PhosphoLingo were similar for the S/T results, with TPs having a peak approximate between .9 and 1.0 (Figure S6G), the TP Tyr phosphosites having a peak approximate between .9 and 1.0 only for PhosBoost (Figure S6G).

While the F1 score assumes equal importance to the precision and recall, depending on the stated objective for the classification

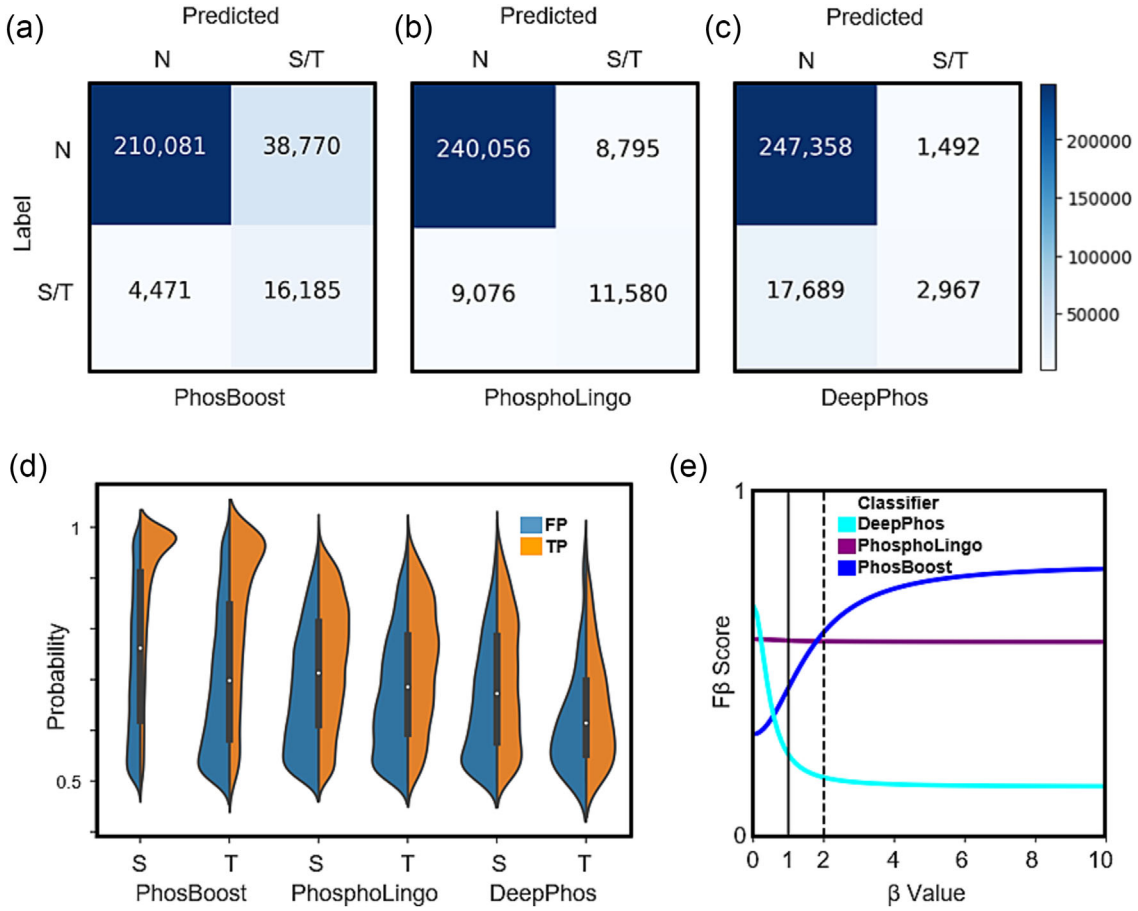

FIGURE 3 Despite lower precision, the PhosBoost S/T model produces more informative predicted probability scores and achieves better performance when recall is prioritized. (a-c) confusion matrices for the S/T model results for PhosBoost, PhosphoLingo, and DeepPhos, respectively (N stands for negative phosphosites). (d) A split violin plot showing the distribution of the predicted probability values for all true positive (TP) and false positive (FP) samples (predicted probability > .5) separated by serines (S) and threonines (T), for PhosBoost, PhosphoLingo, and DeepPhos, respectively. (E) Evaluation of the PhosBoost, PhosphoLingo, and DeepPhos model performances using the Fβ measure at different β values. All results are based on models trained on the *A. thaliana* qPTMplants dataset. The heatmap legend shown is shared across the confusion matrices.

method, it is possible to assess the precision-recall tradeoff using a weighted Fβ score. The generalized Fβ-score produces an F-score for different β values that evaluate the classifier performance under the assumption that recall is β-times as important as precision. To compare the performance of PhosBoost, PhosphoLingo, and DeepPhos, we plotted the Fβ score over different β values for the *A. thaliana* qPTMplants results. Based on this plot we showed that as the β increases, the Fβ score for PhosBoost increases, remains relatively uniform for PhosphoLingo, and decreases for DeepPhos (Figure 3e). A similar observation was made for the Tyr results, whereas the β increases, the Fβ score for PhosBoost increases (Figure S5C). Furthermore, while the F1 score of PhosphoLingo is higher than PhosBoost and DeepPhos (Table 1), when the β is equal to 1.8 the Fβ score of PhosBoost reaches the Fβ-score of PhosphoLingo, suggesting that if recall is considered to be approximately twice as important as precision, PhosBoost performs better than PhosphoLingo and DeepPhos (Figure 3e). Similar results were obtained when comparing the performance of PhosBoost, PhosphoLingo, and DeepPhos using the Fβ score at different β values on the Ramsamy22 dataset, but with a slightly higher β value of 2 for the S/T model (Figure S6H) and a lower β value of 1.6 for the Y model (Figure S6I).

2.6 | A DIAMOND pairwise sequence alignment-based approach can be used to supplement protein phosphorylation prediction methods to improve phosphosite prediction and annotation

Sequence alignment-based approaches were shown to be useful for phosphosite annotation and prediction by searching for similarity between sequences surrounding phosphosites and experimentally derived protein phosphorylation data (Chaudhuri et al., 2015; Maathuis, 2008; Tan et al., 2009). Due to the higher number of predicted false positives by PhosBoost, compared to DeepPhos and PhosphoLingo, we considered using a DIAMOND protein pairwise alignment-based approach to both evaluate the predicted false

positives and improve phosphosite annotation (Buchfink et al., 2021). We first assessed the performance of such an approach when used as a protein phosphorylation prediction method on the *A. thaliana* qPTMplants data. Using DIAMOND, we extracted peptide sequences of size 31 centered at the Ser/Thr phosphosites in the test dataset and conducted pairwise alignments to identify matching experimental phosphosites in the training and validation datasets. We were able to correctly classify 8,852 of the 20,656 experimentally derived Ser and Thr phosphosites, achieving a precision score of .33 and a recall score of .43 (Figure 4a). Furthermore, we found that 17,847 Ser/Thr negative phosphosites matched experimentally derived phosphosites (Figure 4a). For the experimentally derived Tyr phosphosites, we were able to correctly classify 168 of the 654 sites, achieving a precision score of .06 and recall score of .26, and found 2,444 Tyr negative phosphosites that matched experimentally derived phosphosites (Figure S7A).

Next, we used the DIAMOND-based pairwise alignment step to estimate the improvement of the phosphosite annotation, focusing on the predicted false positive sites. We aligned all false positive Ser and Thr phosphosites predicted by PhosBoost, PhosphoLingo, and DeepPhos to the complete qPTMplants database. Phosphosites matching any experimentally derived phosphosite, excluding self-matches, were relabeled as inferred positives (IP) and annotated with the supporting information (Figure 4b). Based on this analysis, we were able to relabel 7,543 FP Ser and Thr phosphosites predicted by PhosBoost, 4,082 by PhosphoLingo, and 482 by DeepPhos (Figure 4c). Thus, in the case of PhosBoost, by combining the inferred positives with the true positive labels, we were able to increase the number of predicted

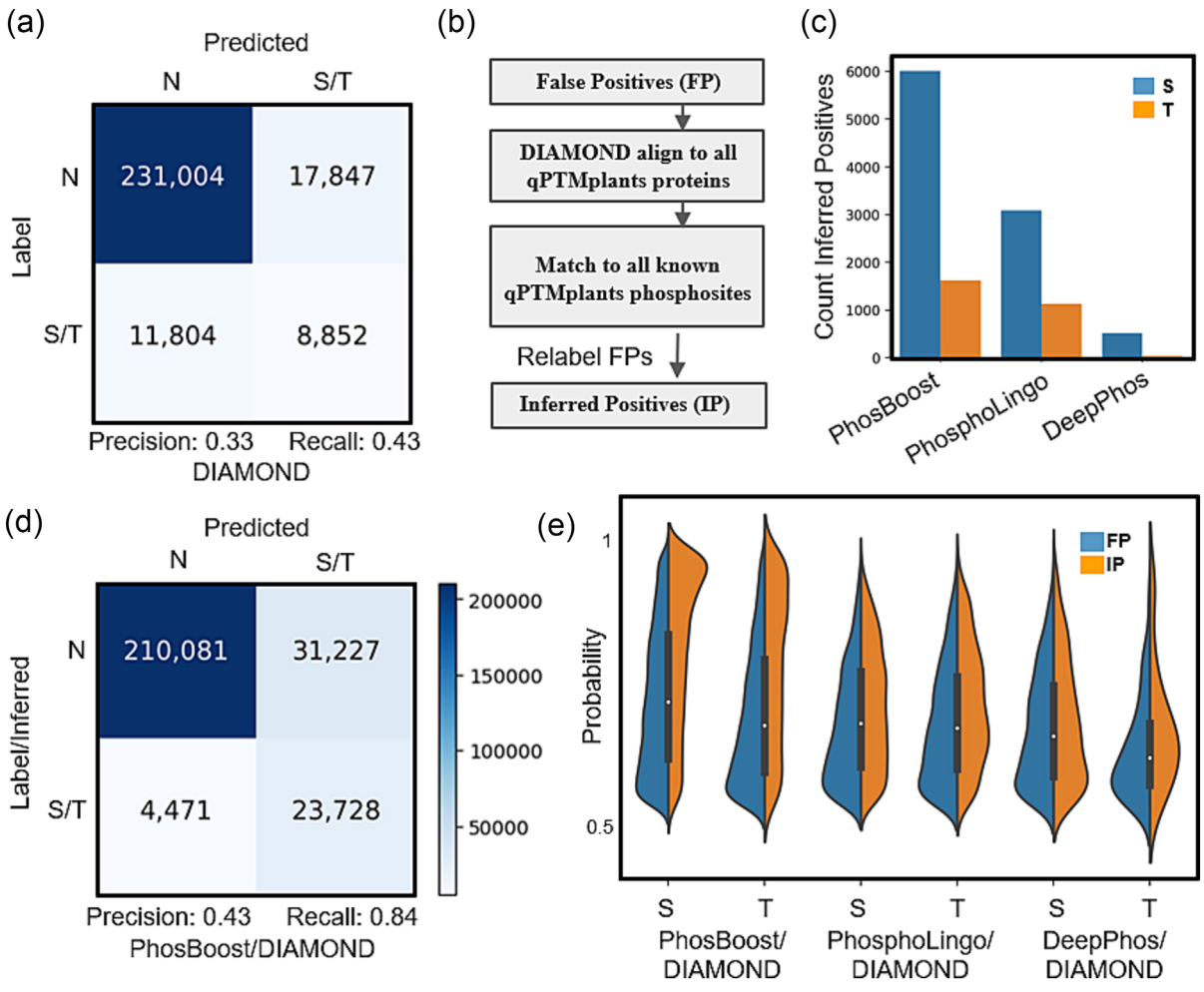

FIGURE 4 Using a DIAMOND-based pairwise alignment analysis improves phosphosite annotation and reduces false positive label uncertainty. (a) A confusion matrix using a DIAMOND-based binary protein phosphorylation prediction was trained and tested on the *A. thaliana* qPTMplants dataset. (b) A schema representing the workflow that uses DIAMOND to evaluate false positive (FP) sites to identify inferred positive (IP) sites. (c) A bar graph showing the number of IP serines (S) and threonines (T) inferred from the PhosBoost, PhosphoLingo, and DeepPhos FP results. (d) Confusion matrix for the PhosBoost S/T model results after combining true positive and IP phosphosites accounting for the FP phosphosites. (e) A split violin plot showing the distribution of the predicted probability values for all FP and IP phosphosites (predicted probability > .5) for the PhosBoost, PhosphoLingo, and DeepPhos results, separated by serines (S) and threonines (T). In all confusion matrices, N stands for non-phosphorylated. The heatmap legend shown is shared across the confusion matrices.

and inferred true positive Ser and Thr phosphosites from 16,185 to 23,728 and reduce the number of false positives from 38,770 to 31,227, while achieving a precision score of .43 and a recall score of .84 (Figure 4d). For the Tyr results, we were able to increase the number of predicted and inferred true positive phosphosites from 391 to 1,375 and reduce the number of false positives from 12,448 to 11,464, while achieving a precision score of .11 and a recall score of .84 (Figure S5A, S7B). We also assessed whether the predicted probability scores for the inferred positive Ser and Thr phosphosites differ from the false positives in PhosBoost, PhosphoLingo, and DeepPhos. When comparing the distribution of the probability scores between the false positives and inferred positives, we observed that the largest difference between the probability distributions, for both Ser and Thr phosphosites, was for PhosBoost, with a distribution peak approximately between probability scores of .9 and 1.0, while the distribution for the predicted probability scores for the false positives and inferred positives were relatively similar for PhosphoLingo and DeepPhos (Figure 4e). In contrast, for the Tyr results, we found that the distribution peaks for both false positives and inferred positives are approximately between probability scores of .5 and .7 (Figure S7C).

2.7 | Assessing the transferability of models trained on *A. thaliana* data at predicting protein phosphorylation in other plant species

We wanted to assess the ability of the S/T and Y PhosBoost models trained on the *A. thaliana* qPTMplants dataset to correctly predict known functional phosphosites in more distantly related plant species such as *Z. mays* and *T. aestivum*. Compared to the available experimental protein phosphorylation for *A. thaliana* in the qPTMplants database, a much smaller amount of data is available for other plant species, including *Z. mays* and *T. aestivum* (Figure S1A). The resulting lack of phosphosite saturation and higher label imbalance in *Z. mays* and *T. aestivum*, compared to *A. thaliana*, suggests that the use of standard metrics, such as precision, recall, AUPRC, and F1 scores, to assess the performance of protein phosphorylation models are not as informative. Therefore, to assess the transferability of models trained on *A. thaliana* data, we used split violin plots to show the distributions of the probability score predicted by the S/T and Y PhosBoost models, across the positive and negative Ser, Thr, and Tyr phosphosites (Figure 5a). The observed difference in the predicted probability scores for the positive and negative Ser, Thr, and Tyr shows that in general, positive phosphosites were predicted with a higher probability score than negative ones, with a peak distribution approximately between .9 and 1, albeit a smaller peak for Thr predictions (Figure 5a). Furthermore, we detected over 40,000, close to 20,000, and close to 10,000 inferred positive Ser, Thr, and Tyr phosphosites, respectively, in both *Z. mays* and *T. aestivum* (Figure 5b).

For a concise prediction comparison, we compiled a short list of functionally important phosphosites, validated in *A. thaliana*, involved in the brassinosteroid (BR), ethylene (ET), and abscisic acid (ABA) pathways. In the BR pathway, we included BRASSINOSTEROID INSENSITIVE 1 (BRI1) Y831 and Y956 (Bojar et al., 2014; Oh et al., 2009), BRASSINOSTEROID-INSENSITIVE2 (BIN2) S187, S203, and Y200 (Kim et al., 2009; Xiong et al., 2017), BRASSINAZOLE RESISTANT1 (BZR1) Thr173 and Thr177 (Ryu et al., 2007), and BRI1-SUPPRESSOR1 (BSU1) Ser251 and Ser764 (Park et al., 2022). In the ET pathway, we included ETHYLENE INSENSITIVE3 (EIN3) Thr174 and Thr592 (Zhao & Guo, 2011) and ACC SYNTHASE 6 Ser480, Ser483, and Ser488 (Liu & Zhang, 2004). In the ABA pathway, we included OPEN STOMATA1 (OST1) Ser175 (Belin et al., 2006), SLOW ANION CHANNEL-ASSOCIATED1 (SLAC1) Ser59 and Ser120 (Brandt et al., 2015), and ABSCISIC ACID INSENSITIVE5 (ABI5) Ser42, Ser145, and Thr201 (Y. Wang et al., 2013). The top blast hit in *Z. mays* and *T. aestivum* for each of the *A. thaliana* protein sequences was used to identify the reciprocal phosphosite. We then used PhosBoost and PhosphoLingo, trained on the *A. thaliana* qPTMplants dataset to generate the predicted probability scores for each phosphosite (Figure 5c, Table S2). The prediction results show that except for BRI1-Y956, BIN2-S187, BSU1-S251, and EIN3-T174/T592 for PhosBoost and PhosphoLingo, and BRI1-Y831, BSU1-S764, and ABI5-T201 for PhosphoLingo, most of *the A. thaliana* sites were correctly predicted by both methods (Figure 5c). For the *Z. mays* and *T. aestivum*, PhosBoost correctly predicted more of the phosphosites than PhosphoLingo and with a higher predicted probability score for most phosphosites (Figure 5c). PhosBoost also generated more consistent predictions across the three species tested, with the exception of EIN3-T174 and ABI5-T201 for PhosBoost and PhosphoLingo, and BZR1-S173, BSU1-S764, ACS6-S483/S488, and ABI5-S42/T201 (Figure 5c).

2.8 | Using PhosBoost for genome-wide protein phosphorylation predictions and integration within genome browsers

In this study, we aimed to use PhosBoost as a scalable machine-learning method for generating genome-wide protein phosphorylation predictions. For this, we trained new S/T and Y PhosBoost models on the complete qPTMplants datasets to be used for plant phosphosite predictions. We then used PhosBoost to conduct genome-wide protein phosphorylation predictions in four plant species (one accession per species): wheat (Chinese Spring), oat (Sang), barley (Morex), and maize (B73), using one representative protein sequence for each gene model. Additionally, the DIAMOND pairwise alignment analysis was used to improve the annotation of all phosphosites using the complete qPTMplants dataset as a reference. To achieve this, we developed a straightforward approach that converts the prediction and annotation results into a GFF3 format, allowing for direct integration within JBrowse genome browsers (Figure 6a) (Diesh et al., 2023). In this example, the labels of phosphosites with predicted probability above .9 were marked in bold font, and labels were color-coded as follows: blue for phosphosites inferred by DIAMOND pairwise sequence alignment, red for phosphosites with predicted probability scores above .5, pink if both cases apply (Figure 6a). Each

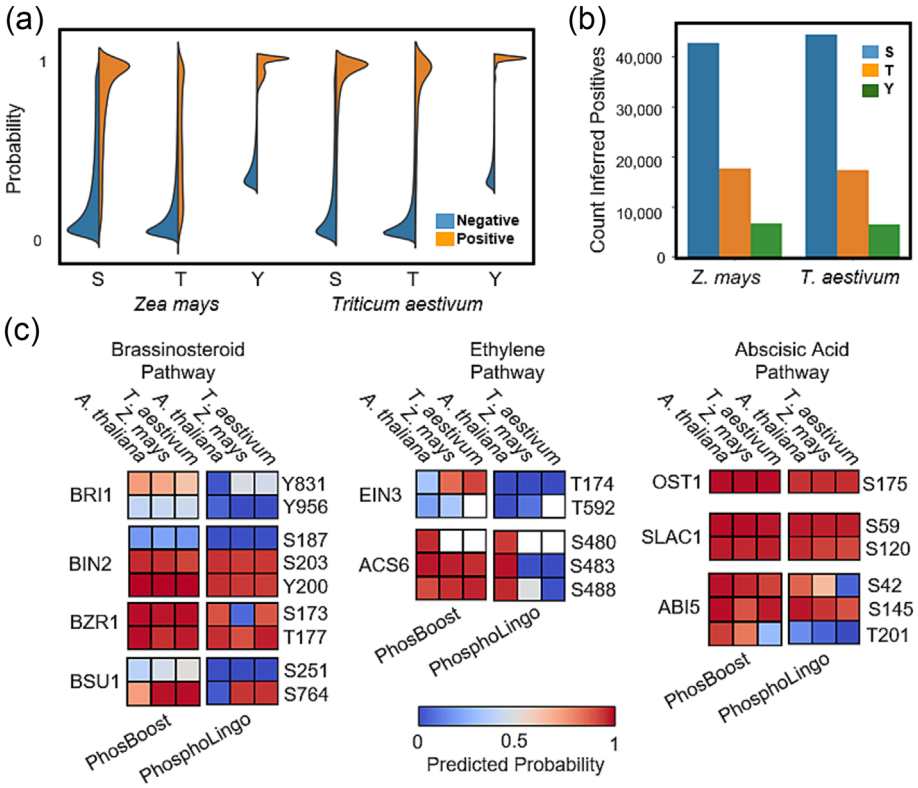

FIGURE 5 PhosBoost protein phosphorylation predictions are transferable across plant species and perform better than PhosphoLingo on a number of functionally important phosphosites. (a) A split violin plot showing the distribution of the predicted probability scores for all phosphosites present in the *Zea mays* and *Triticum aestivum* qPTMplants database predicted using the S/T and Y PhosBoost models trained on the *A. thaliana* qPTMplants dataset. The results show the probability scores for all negative phosphosites (blue) and positive phosphosites (orange) separated by serines (S), threonines (T), and tyrosines (Y) phosphosites. (b) A bar graph showing the number of inferred positive (IP) S, T, and Y phosphosites as detected by using DIAMOND to align the predicted false positive phosphosites by the S/T and Y PhosBoost models to the complete qPTMplants database. (c) Analysis of the predicted probability scores for PhosBoost and PhosphoLingo on a small number of verified functional *A. thaliana* phosphosites in the brassinosteroid, ethylene, and abscisic acid phytohormone pathways and the matching phosphosites in the top blast hit in *Z. mays* and *T. aestivum*. Non-conserved sites are filled with white background.

phosphosite contains additional metadata such as the phosphosite ID, predicted probability score and DIAMOND pairwise sequence alignment matches (Figure 6b). These results are now accessible through the GrainGenes (Yao et al., 2022) and MaizeGDB databases (Woodhouse et al., 2021).

3 | DISCUSSION

Protein phosphorylation is an important and dynamic post-translational modification that has a wide-spread effect on many biological processes by regulating protein functions and interactions (Nishi et al., 2014). Compared to *A. thaliana*, experimental phosphorylation data for most other plant species are either not available or are relatively small and would require additional high-throughput phosphoproteomic experiments to achieve a similar degree of saturation (Figure S1A-C). Therefore, the prediction of protein phosphosites is instrumental in elucidating possible regulatory phosphosites where extensive experimental phosphoproteomic data is not available (Meng et al., 2022). While the use of state-of-the-art deep learning and pLMs for protein phosphorylation prediction improved performance over existing methods (Zuallaert et al., 2022), classical machine learning approaches, such as gradient boosting trees, remain as powerful and scalable classification methods (Anghel et al., 2019; Lyashevska et al., 2021). Furthermore, ensemble methods that combine different machine-learning classification methods can improve classification performance by learning from multiple weak classifiers (Dietterich, 1997). Choosing appropriate class weights during training can have a substantial effect on the performance of the trained machine-learning classification model, including the popular use of balanced class weights for highly imbalanced data (Cui et al., 2019; Liu & Zhou, 2006). In designing PhosBoost, we observed that a stacking classifier consisting of two CatBoost base classifiers, one trained with balanced class weights and one trained with equal class weights, outperforms both independently trained CatBoost classifiers on two different datasets, providing an increased recall with no cost to precision for both the S/T and Y models (Figure 1c-h, Figure S3A-F). We hypothesize that the use of balanced class weights to train one base classifier helps address the class imbalance by giving more weight to the underrepresented positive class while

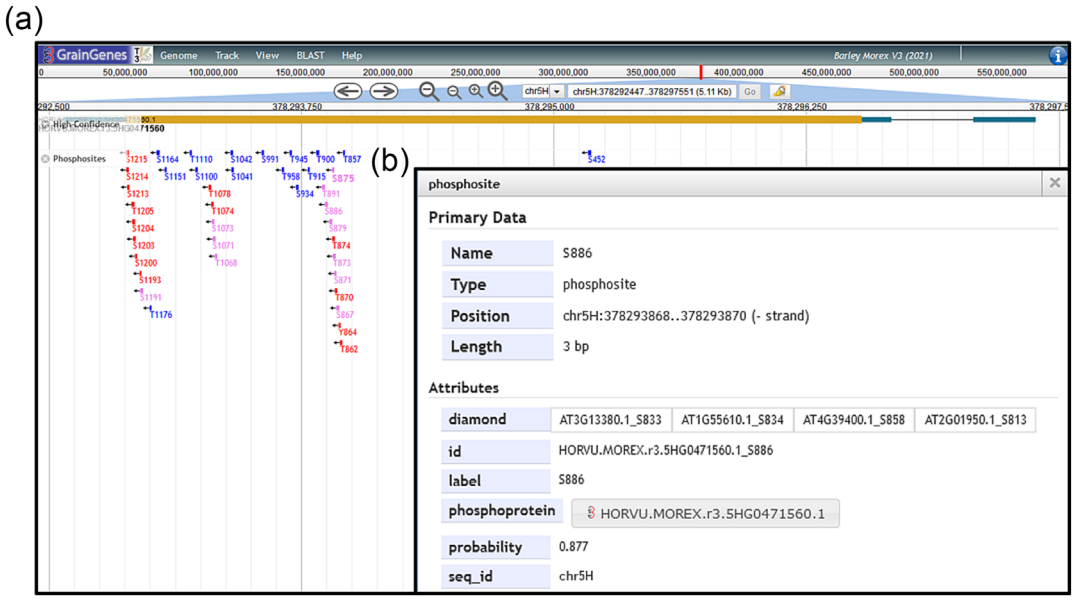

FIGURE 6 Integrating the PhosBoost prediction results as a track within the JBrowse genome browser for better accessibility. For this example, we generated a GFF3 file that contains all positive and inferred predicted phosphosites by PhosBoost for all protein sequences in the *H. vulgare* Morex version 3 genome assembly, using one representative protein sequence for each gene model. All phosphosites were mapped to their respective 3 bp genomic coordinates. (a) Screenshot of the PhosBoost predicted phosphosites track view for the gene HORVU.MOREX.r3.5HG0471560.1. In this example, phosphosites were color-coded as follows: blue for phosphosites inferred by DIAMOND pairwise sequence alignment, red for phosphosites with predicted probability scores above .5, and pink if both cases apply. The labels of phosphosites with predicted probability scores above .9 are shown in bold font. (b) A view of one of the PhosBoost predicted phosphosites, showing the phosphosite ID, predicted probability score, and the ID of the DIAMOND pairwise sequence alignment matches with experimentally derived phosphosites in the qPTMplants database.

the base classifier trained with equal class weights captures the overall characteristics of the dataset, effectively learning from the majority negative class, allowing the stacking classifier to synergistically learn from these complementary approaches.

Next, we benchmarked the performance of PhosBoost and two existing protein phosphorylation prediction methods, namely PhosphoLingo and DeepPhos, on two separate datasets. For the S/T model, we observed that the performance of PhosBoost, based on the AUPRC scores, was comparable to PhosphoLingo on the *A. thaliana* qPTMplants, lower on the Ramasamy22 data, and higher than DeepPhos on both datasets (Figure 2b, S4B, Table 1). On the other hand, for the Y model, PhosBoost performed better than PhosphoLingo and DeepPhos on both datasets (Figure 2e, S4E, Table 1). Different data-centric factors, such as dataset size and label imbalance, can have a substantial impact on model performance (L. Dou et al., 2021; Lyashevska et al., 2021). A comparison of the S/T and Y model performances in relation to data size and imbalance, suggests that data size impacted the S/T model performance and data imbalance impacted the Y model performance when comparing the models trained on the *A. thaliana* qPTMplants and Ramasamy22 datasets (Figure S2A-C, 2a-f, S4A-F). Based on our results, we hypothesize that the performance of PhosBoost improves when the input data increases in size, as observed for the S/T model (Figure S2A-B), or increases in label imbalance, as observed for the Y model (Figure S2C). We also demonstrate that PhosBoost consistently achieves higher recall scores at all probability thresholds compared to PhosphoLingo and DeepPhos (Figure 2c, f, S4C, S4F). Thus, we show that PhosBoost can be competitive with existing protein phosphorylation prediction methods, particularly when higher recall and improved predicted phosphosite coverage are beneficial.

Despite the improved recall score, PhosBoost presents a precision-recall tradeoff associated with an increased number of predicted false positives compared to PhosphoLingo and DeepPhos (Figure 3a-c, Figure S5A). We assessed the performance of PhosBoost, PhosphoLingo, and DeepPhos under the assumption that, depending on the aim of the prediction method, the importance of recall can outweigh precision. We show that PhosBoost has a higher Fβ score when the value of the β parameter is approximately 2 for both the *A. thaliana* qPTMplants and Ramsamy22 datasets, in the S/T and Y models (Figure 3d, S6H-I). Therefore, in view of the precision-recall tradeoff, we suggest that when a recall is prioritized and considered to be twice as important as precision, PhosBoost achieves higher performance than PhosphoLingo and DeepPhos on both the *A. thaliana* qPTMplants and Ramasamy22 datasets. Furthermore, we show that PhosBoost consistently produces predicted probability scores that are more indicative of whether a phosphosite are true or false positive, providing a more confident and informative predicted probability score (Figure 3d, S6G). Taken together, our results suggest that PhosBoost achieves higher performance than PhosphoLingo and DeepPhos when recall is prioritized over precision while providing more confidence through more informative predicted probability scores.

Due to complex phosphorylation and dephosphorylation dynamics and technical limitations, many phosphosites remain undetected by high throughput phosphoproteomic approaches (Doll & Burlingame, 2015; Gelens & Saurin, 2018; Iakoucheva, 2004). One of the approaches used to infer if a phosphosite is positive despite the lack of experimental data is through the identification of conserved phosphosites with experimental evidence (Chaudhuri et al., 2015). To partially address the problem of negative phosphosite label uncertainty, we have developed a supplementary pairwise sequence alignment analysis step using DIAMOND to improve phosphosite annotation. We show that this approach can be used by any protein phosphorylation classification method to increase the number of inferred positive phosphosites (Figure 4c), reduce the number of false positives (Figure 4d), and provide useful information about the matching experimental phosphosites. Furthermore, by comparing the probability score distribution of the false and inferred positive phosphosites, we show that PhosBoost produces higher predicted probability scores for inferred positives than false positives for Ser and Thr phosphosites, providing a more confident and informative predicted probability score than PhosphoLingo and DeepPhos (Figure 4e). For Tyr phosphosites, PhosBoost was not as informative at predicting inferred positives compared to false positives (Figure S7C), possibly due to higher label imbalance (Figure S2C). Label uncertainty and mislabeling is a common problem for protein phosphorylation classification methods that could be partially addressed by improving phosphosite annotation prior to training and testing, with a possible benefit of improved model performance and interoperability.

One of our goals when developing PhosBoost was to develop a method that is scalable for genome-wide protein phosphorylation prediction. Because a large portion of the experimental protein phosphorylation data in the qPTMplants belongs to *A. thaliana* phosphosites, we assessed whether PhosBoost trained on the *A. thaliana* dataset can effectively predict phosphosites in other plant species. Comparison of the distribution of the predicted probability scores for Ser, Thr, and Tyr, provides compelling evidence that the S/T and Y models can provide informative predicted probability scores that effectively distinguish between true and false positive phosphosites. We also show that the DIAMOND pairwise alignment step can be used to improve the phosphosite annotation of a large number of Ser, Thr, and Tyr phosphosites, in both *Z. Mays* and *T. aestivum*. A comparative analysis of PhosBoost and PhosphoLingo on a small number of known functional phosphosites from *A. thaliana* with matching phosphosites from *Z. mays* and *T. aestivum* showed that PhosBoost correctly predicted a larger number of phosphosites with higher predicted probability scores than PhosphoLingo. Based on these results, we trained final the final PhosBoost S/T and Y models on the complete qPTMplants protein phosphorylation dataset and conducted genome-wide protein phosphorylation prediction on four plant species, *Z. mays*, *T. aestivum*, *Avena sativa*, and *Hordeum vulgare*. To facilitate the accessibility to the prediction data, we developed a method that converts the prediction results and the pairwise DIAMOND alignment results into a file that can be directly incorporated and visualized on the genome browser (Figure 6a-b).

4 | CONCLUSION

We present PhosBoost, a general protein phosphorylation prediction method that uses pLMs and a stacking classifier composed of CatBoost gradient boosting tree base classifiers. We show evidence to suggest that PhosBoost has improved performance when working with large and imbalanced datasets, showing comparable results for Ser/Thr classification and improved Tyr classification. We show that PhosBoost consistently achieves higher recall and more informative predicted probability scores, making PhosBoost particularly useful when recall is prioritized over precision and a higher coverage of predicted phosphosites is beneficial. We developed a DIAMOND pairwise sequence alignment analysis step to reduce phosphosite label uncertainty and improve phosphosite annotation. We show that PhosBoost is scalable to genome-wide protein phosphorylation predictions and implement a straightforward method for integrating prediction and annotation results directly in the genome browser to facilitate accessibility.

5 | MATERIALS AND METHODS

5.1 | Data sources and generation of input embedding data

The complete list of phosphorylated residues from the qPTMplants database for protein post-translational modification was used (Xue et al., 2022). Because the qPTMplants database does not provide the protein sequences for the annotated phosphoproteins, protein sequences were manually obtained from genomic databases such as UniProt (The UniProt Consortium et al., 2023), Phytozome (Goodstein et al., 2012), and EnsemblPlants (Yates et al., 2022), and processed with BioPython (v.1.81) (Cock et al., 2009). Proteins that are not present in genomic databases or contain missing or non-canonical amino-acid sequences were discarded. The phosphorylation positions were cross-referenced to ensure that the qPTMplants sites matched the protein sequences, and no-matching sequences were removed. Protein sequences containing non-canonical or missing amino acids were removed. The resulting FASTA file was used directly as the input for calculating protein and amino acid embedding vectors using the pretrained ProtT5-XL-U50 pLM (Elnaggar et al., 2021). The Python code used to generate the embeddings was obtained from the ProtTrans GitHub repository (https://github.com/agemagician/ProtTrans) and is based on the ProtT5 pLM architecture. The code was modified to specifically return the amino-acid embedding vectors of S/T/Y residues and the protein-wise average embedding vector (available at https://github.com/eporetsky/PhosBoost). The combined S/T/Y amino-acid and protein-wise average embedding data were used, without modification, as input data for the PhosBoost classifiers. Using the generated amino acid embedding vectors and the average protein embedding vectors, all sites included in the qPTMplants database were extracted as a positive phosphorylated site set, and all the remaining S/T/Y sites were extracted as a negative phosphorylation

set. Initial training and evaluation on predictive classification models were conducted using the data obtained from the complete *A. thaliana* phosphorylation dataset. For the training and validation of PhosBoost, only the *A. thaliana* data from the qPTMplants database was used. The complete *A. thaliana* dataset was randomly split into stratified training, validation, and testing sets using a 60%–20%–20%, split, respectively.

5.2 | Overview of the PhosBoost protein phosphorylation classifier

A stacking classifier is a two-step ensemble learning method composed of an initial stack of base classifiers followed by a final meta-classifier that integrates the results of the base classifiers. In the first step, the training data is used to separately train the base classifiers specified within the stack. In the second step, the predicted probability scores from each base classifier are used as new training data features for training the final meta-classifier. The metaclassifier is trained using a cross-validated prediction approach to reduce over-fitting by not training on the same data used to train the base classifiers. Stacking classifiers have been shown to often perform better than the individual base classifiers by combining different types of classifiers that may have complementary strengths and weaknesses while reducing bias and variance. We implemented PhosBoost as a stacking classifier composed of two stacked base-classifiers and a metaclassifier, using a 5-fold cross-validation approach. The stacking classifier is comprised of two CatBoost base classifiers (v1.1.1) (Dorogush et al., 2018), one trained using the balanced class weight parameter to modify class weights according to label frequencies (auto_class_weights="balanced"), and one trained using the default equal class weights parameters. The predicted probability scores produced by each of the two base-classifiers were then used as input training data features for the logistic regression metaclassifier, using the balanced class weight parameter. The stacking classifier and logistic-regression metaclassifier were implemented using the Python scikit-learn package (v1.2.1) (Pedregosa et al., 2011). The CatBoost API is compatible with the scikit-learn architecture, enabling direct integration of the base-classifiers within the stacking classifier. Furthermore, to provide support for the improved performance of the stacking classifier, the stacking PhosBoost classifier performance was compared to the performance of the two independently trained CatBoost classifiers outside of the stacking classifier. Hyper-parameter tuning was conducted on the training and validation sets, using the Bayesian optimization function BayesianOptimization from the bayes_opt Python package (v1.4.2) (Nogueira, 2014). The Bayesian optimization method was used to fine-tune the following hyper-parameters: "n_estimators" (between 50 and 2000), "depth" (between 2 and 10), and "learning_rate" (between .05 and .5) by optimizing the F1-score metric over 100 iterations. Hyper-parameters optimization was conducted separately for the two independent CatBoost classifiers, one trained with equal class weights and one trained with balanced class weights, and separately for the ST and Y models. The obtained model parameters were then used by the two respective CatBoost base-classifiers within the ST and Y model PhosBoost stacking classifiers. The "n_estimators" and "depth" hyper-parameters were rounded to the nearest integer, as instructed by the bayes_opt Python package (v1.4.2). The obtained values for the hyper-parameters are available (Table S1).

5.3 | Evaluation of protein phosphorylation predictive model performances

We benchmarked the performance of PhosBoost compared to two established protein phosphorylation classification methods, namely PhosphoLingo (v0.1.0) (Zuallaert et al., 2022) and DeepPhos (Luo et al., 2019). All three classification methods produce two separate binary classifiers, the S/T and Y models. PhosBoost and PhosphoLingo used the same training and validation sets for hyperparameter tuning, while DeepPhos used the combined training and validation sets for internal hyperparameter tuning and training. For PhosphoLingo we used the pre-trained ProtT5-XL-U50 pLM model to train a protein phosphorylation model under the "full" setting. The default settings were used in the training and testing process of DeepPhos.

5.4 | Building a DIAMOND-based inference method to improve phosphosite annotations

To enhance phosphosite annotation, we employed a sequence-based DIAMOND alignment step (v2.1.6) (Buchfink et al., 2021). We extracted a peptide of length 31 for each candidate phosphosite of Ser, Thr, and Tyr within a protein sequence, with the phosphosite at the center of the sequence. For phosphosites positioned at the edges of protein sequences, a peptide of length 31 was extracted but with the phosphosite position being determined based on the distance from either edge of the protein sequence. After obtaining all peptide sequences, each peptide was aligned using DIAMOND (−-masking none --ultra-sensitive --max-target-seqs 100) against all protein sequences in the complete qPTMplants database. All pairwise sequence alignment results were tested for each query to identify matches with experimentally derived phosphosites. All query phosphosites with matches to the qPTMplants database were labeled as inferred positives, and the matches were compiled as a list to improve phosphosite annotation and provide additional supportive information.

5.5 | Data and code availability

A detailed markdown page provides explanations for all analysis steps, code for reproducing results and figures, and links to raw data and results, which are available on GitHub at https://github.com/eporetsky/PhosBoost. The protein phosphorylation data used in this study is available for download directly from the qPTMplants database (Xue et al., 2022) and PhosphoLingo GitHub repository (Zuallaert

et al., 2022). Additionally, we assembled several helper functions to facilitate raw data processing and conversion between different file formats used by different protein phosphorylation prediction methods as a python package named PTMtools that is available through the official PyPI repository.

AUTHOR CONTRIBUTIONS

EP, CMA, and TZS designed the research. EP performed the research and drafted the manuscript. EP, CMA, and TZS revised and finalized the manuscript.

ACKNOWLEDGMENTS

This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D. USDA is an equal opportunity provider and employer.

CONFLICT OF INTEREST STATEMENT

The Authors did not report any conflict of interest.

ORCID

Elly Poretsky https://orcid.org/0000-0002-6116-6827

Taner Z. Sen https://orcid.org/0000-0002-5553-6190

REFERENCES

Álvarez-Salamero, C., Castillo-González, R., & Navarro, M. N. (2017). Lighting up T lymphocyte signaling with quantitative Phosphoproteomics. *Frontiers in Immunology*, *8*, 938. https://doi.org/10.3389/fimmu.2017.00938

Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G., & Pandey, A. (2007). A curated compendium of phosphorylation motifs. *Nature Biotechnology*, *25*(3), 285–286. https://doi.org/10.1038/nbt0307-285

Anghel, A., Papandreou, N., Parnell, T., De Palma, A., & Pozidis, H. (2019). *Benchmarking and optimization of gradient boosting decision tree algorithms* (arXiv:1809.04559). arXiv. http://arxiv.org/abs/1809.04559

Ardito, F., Giuliani, M., Perrone, D., Troiano, G., & Muzio, L. L. (2017). The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (review). *International Journal of Molecular Medicine*, *40*(2), 271–280. https://doi.org/10.3892/ijmm.2017.3036

Belin, C., De Franco, P.-O., Bourbousse, C., Chaignepain, S., Schmitter, J.-M., Vavasseur, A., Giraudat, J., Barbier-Brygoo, H., & Thomine, S. (2006). Identification of features regulating OST1 kinase activity and OST1 function in guard cells. *Plant Physiology*, *141*(4), 1316–1327. https://doi.org/10.1104/pp.106.079327

Bojar, D., Martinez, J., Santiago, J., Rybin, V., Bayliss, R., & Hothorn, M. (2014). Crystal structures of the phosphorylated BRI 1 kinase domain and implications for brassinosteroid signal initiation. *The Plant Journal*, *78*(1), 31–43. https://doi.org/10.1111/tpj.12445

Bordin, N., Dallago, C., Heinzinger, M., Kim, S., Littmann, M., Rauer, C., Steinegger, M., Rost, B., & Orengo, C. (2023). Novel machine learning approaches revolutionize protein knowledge. *Trends in Biochemical Sciences*, *48*(4), 345–359. https://doi.org/10.1016/j.tibs.2022.11.001

Brandt, B., Munemasa, S., Wang, C., Nguyen, D., Yong, T., Yang, P. G., Poretsky, E., Belknap, T. F., Waadt, R., Alemán, F., & Schroeder, J. I. (2015). Calcium specificity signaling mechanisms in abscisic acid signal transduction in Arabidopsis guard cells. *eLife*, *4*, e03599. https://doi.org/10.7554/eLife.03599

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Chaudhuri, R., Sadrieh, A., Hoffman, N. J., Parker, B. L., Humphrey, S. J., Stöckli, J., Hill, A. P., James, D. E., & Yang, J. Y. H. (2015). PhosphOrtholog: A web-based tool for cross-species mapping of orthologous protein post-translational modifications. *BMC Genomics*, *16*(1), 617. https://doi.org/10.1186/s12864-015-1820-x

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). *Class-balanced loss based on effective number of samples (arXiv:1901.05555)*. arXiv. http://arxiv.org/abs/1901.05555

Diesh, C., Stevens, G. J., Xie, P., De Jesus Martinez, T., Hershberg, E. A., Leung, A., Guo, E., Dider, S., Zhang, J., Bridge, C., Hogue, G., Duncan, A., Morgan, M., Flores, T., Bimber, B. N., Haw, R., Cain, S., Buels, R. M., Stein, L. D., & Holmes, I. H. (2023). JBrowse 2: A modular genome browser with views of synteny and structural variation. *Genome Biology*, *24*(1), 74. https://doi.org/10.1186/s13059-023-02914-z

Dietterich, T. G. (1997). Machine-learning research four current directions. *AI Magazine*, *18*(4), 97–136.

Doll, S., & Burlingame, A. L. (2015). Mass spectrometry-based detection and assignment of protein posttranslational modifications. *ACS Chemical Biology*, *10*(1), 63–71. https://doi.org/10.1021/cb500904b

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: Gradient boosting with categorical features support (arXiv:1810.11363)*. arXiv. http://arxiv.org/abs/1810.11363

Dou, L., Yang, F., Xu, L., & Zou, Q. (2021). A comprehensive review of the imbalance classification of protein post-translational modifications. *Briefings in Bioinformatics*, *22*(5), bbab089. https://doi.org/10.1093/bib/bbab089

Dou, Y., Yao, B., & Zhang, C. (2014). PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids*, *46*(6), 1459–1469. https://doi.org/10.1007/s00726-014-1711-5

Dressano, K., Weckwerth, P. R., Poretsky, E., Takahashi, Y., Villarreal, C., Shen, Z., Schroeder, J. I., Briggs, S. P., & Huffaker, A. (2020). Dynamic regulation of pep-induced immunity through post-translational control of defence transcript splicing. *Nature Plants*, *6*(8), 1008–1019. https://doi.org/10.1038/s41477-020-0724-1

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2021). ProtTrans: Towards cracking the language of Lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *1–1*, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

Friso, G., & Van Wijk, K. J. (2015). Update: Post-translational protein modifications in plant metabolism. *Plant Physiology*, *169*, pp.01378.2015. https://doi.org/10.1104/pp.15.01378

Gao, J., Thelen, J. J., Dunker, A. K., & Xu, D. (2010). Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Molecular & Cellular Proteomics*, *9*(12), 2586–2600. https://doi.org/10.1074/mcp.M110.001388

Gelens, L., & Saurin, A. T. (2018). Exploring the function of dynamic phosphorylation-Dephosphorylation cycles. *Developmental Cell*, *44*(6), 659–663. https://doi.org/10.1016/j.devcel.2018.03.002

Ghelis, T. (2011). Signal processing by protein tyrosine phosphorylation in plants. *Plant Signaling & Behavior*, *6*(7), 942–951. https://doi.org/10.4161/psb.6.7.15261

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Gough, C., & Sadanandom, A. (2021). Understanding and exploiting post-translational modifications for plant disease resistance. *Biomolecules*, *11*(8), 1122. https://doi.org/10.3390/biom11081122

Guo, H., Ahn, H.-K., Sklenar, J., Huang, J., Ma, Y., Ding, P., Menke, F. L. H., & Jones, J. D. G. (2020). Phosphorylation-regulated activation of the Arabidopsis RRS1-R/RPS4 immune receptor complex reveals two distinct effector recognition mechanisms. *Cell Host & Microbe*, *27*(5), 769–781.e6. https://doi.org/10.1016/j.chom.2020.03.008

Humphrey, S. J., James, D. E., & Mann, M. (2015). Protein phosphorylation: A major switch mechanism for metabolic regulation. *Trends in Endocrinology and Metabolism*, *26*(12), 676–687. https://doi.org/10.1016/j.tem.2015.09.013

Iakoucheva, L. M. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research*, *32*(3), 1037–1049. https://doi.org/10.1093/nar/gkh253

Ilzhöfer, D., Heinzinger, M., & Rost, B. (2022). SETH predicts nuances of residue disorder from protein embeddings. *Frontiers in Bioinformatics*, *2*, 1019597. https://doi.org/10.3389/fbinf.2022.1019597

Ismail, H. D., Jones, A., Kim, J. H., Newman, R. H., & Kc, D. B. (2016). RF-Phos: A novel general phosphorylation site prediction tool based on random Forest. *BioMed Research International*, *2016*, 1–12, 3281590. https://doi.org/10.1155/2016/3281590

Jamal, S., Ali, W., Nagpal, P., Grover, A., & Grover, S. (2021). Predicting phosphorylation sites using machine learning by integrating the sequence, structure, and functional information of proteins. *Journal of Translational Medicine*, *19*(1), 218. https://doi.org/10.1186/s12967-021-02851-0

Kim, T.-W., Guan, S., Sun, Y., Deng, Z., Tang, W., Shang, J.-X., Sun, Y., Burlingame, A. L., & Wang, Z.-Y. (2009). Brassinosteroid signal transduction from cell-surface receptor kinases to nuclear transcription factors. *Nature Cell Biology*, *11*(10), 1254–1260. https://doi.org/10.1038/ncb1970

La Fuente, D., Van Bentem, S., & Hirt, H. (2009). Protein tyrosine phosphorylation in plants: More abundant than expected. *Trends in Plant Science*, *14*(2), 71–76. https://doi.org/10.1016/j.tplants.2008.11.003

Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K., & Rost, B. (2021). Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports*, *11*(1), 23916. https://doi.org/10.1038/s41598-021-03431-4

Liu, S., Cui, C., Chen, H., & Liu, T. (2022). Ensemble learning-based feature selection for phosphorylation site detection. *Frontiers in Genetics*, *13*, 984068. https://doi.org/10.3389/fgene.2022.984068

Liu, Y., & Zhang, S. (2004). Phosphorylation of 1-Aminocyclopropane-1-carboxylic acid Synthase by MPK6, a stress-responsive mitogen-activated protein kinase, induces ethylene biosynthesis in Arabidopsis[W]. *The Plant Cell*, *16*(12), 3386–3399. https://doi.org/10.1105/tpc.104.026609

Liu, X.-Y., & Zhou, Z.-H. (2006). The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study. In *Sixth international conference on data mining (ICDM'06)* (pp. 970–974). IEEE. https://doi.org/10.1109/ICDM.2006.158

Luo, F., Wang, M., Liu, Y., Zhao, X.-M., & Li, A. (2019). DeepPhos: Prediction of protein phosphorylation sites with deep learning. *Bioinformatics*, *35*(16), 2766–2773. https://doi.org/10.1093/bioinformatics/bty1051

Lv, H., Dao, F.-Y., Zulfiqar, H., & Lin, H. (2021). DeepIPs: Comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. *Briefings in Bioinformatics*, *22*(6), bbab244. https://doi.org/10.1093/bib/bbab244

Lyashevska, O., Malone, F., MacCarthy, E., Fiehler, J., Buhk, J.-H., & Morris, L. (2021). Class imbalance in gradient boosting classification algorithms: Application to experimental stroke data. *Statistical Methods in Medical Research*, *30*(3), 916–925. https://doi.org/10.1177/0962280220980484

Maathuis, F. J. M. (2008). Conservation of protein phosphorylation sites within gene families and across species. *Plant Signaling & Behavior*, *3*(11), 1011–1013. https://doi.org/10.4161/psb.6721

Maiti, S., Hassan, A., & Mitra, P. (2020). Boosting phosphorylation site prediction with sequence feature-based machine learning. *Proteins: Structure, Function, and Bioinformatics*, *88*(2), 284–291. https://doi.org/10.1002/prot.25801

Meng, L., Chan, W.-S., Huang, L., Liu, L., Chen, X., Zhang, W., Wang, F., Cheng, K., Sun, H., & Wong, K.-C. (2022). Mini-review: Recent advances in post-translational modification site prediction based on deep learning. *Computational and Structural Biotechnology Journal*, *20*, 3522–3532. https://doi.org/10.1016/j.csbj.2022.06.045

Mühlenbeck, H., Bender, K. W., & Zipfel, C. (2021). Importance of tyrosine phosphorylation for transmembrane signaling in plants. *Biochemical Journal*, *478*(14), 2759–2774. https://doi.org/10.1042/BCJ20210202

Nishi, H., Shaytan, A., & Panchenko, A. R. (2014). Physicochemical mechanisms of protein regulation by phosphorylation. *Frontiers in Genetics*, *5*, 270. https://doi.org/10.3389/fgene.2014.00270

Nogueira, F. (2014). *Bayesian optimization: Open source constrained global optimization tool for python*. https://github.com/fmfn/BayesianOptimization

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. *Computational and Structural Biotechnology Journal*, *19*, 1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022

Oh, M.-H., Wang, X., Kota, U., Goshe, M. B., Clouse, S. D., & Huber, S. C. (2009). Tyrosine phosphorylation of the BRI1 receptor kinase emerges as a component of brassinosteroid signaling in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, *106*(2), 658–663. https://doi.org/10.1073/pnas.0810249106

Park, C. H., Bi, Y., Youn, J.-H., Kim, S.-H., Kim, J.-G., Xu, N. Y., Shrestha, R., Burlingame, A. L., Xu, S.-L., Mudgett, M. B., Kim, S.-K., Kim, T.-W., & Wang, Z.-Y. (2022). Deconvoluting signals downstream of growth and immune receptor kinases by phosphocodes of the BSU1 family phosphatases. *Nature Plants*, *8*(6), 646–655. https://doi.org/10.1038/s41477-022-01167-1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pejaver, V., Hsu, W.-L., Xin, F., Dunker, A. K., Uversky, V. N., & Radivojac, P. (2014). The structural and functional signatures of proteins that undergo multiple events of post-translational modification: Structural and functional signatures of PTM crosstalk. *Protein Science*, *23*(8), 1077–1093. https://doi.org/10.1002/pro.2494

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc.. https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million

protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

Ryu, H., Kim, K., Cho, H., Park, J., Choe, S., & Hwang, I. (2007). Nucleocytoplasmic shuttling of BZR1 mediated by phosphorylation is essential in *Arabidopsis* Brassinosteroid signaling. *The Plant Cell*, *19*(9), 2749–2762. https://doi.org/10.1105/tpc.107.053728

Silva-Sanchez, C., Li, H., & Chen, S. (2015). Recent advances and challenges in plant phosphoproteomics. *Proteomics*, *15*(5–6), 1127–1141. https://doi.org/10.1002/pmic.201400410

Tan, C. S. H., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M. O., Jørgensen, C., Bader, G. D., Aebersold, R., Pawson, T., & Linding, R. (2009). Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Science Signaling*, *2*(81), ra39. https://doi.org/10.1126/scisignal.2000316

Thapa, N., Chaudhari, M., Iannetta, A. A., White, C., Roy, K., Newman, R. H., Hicks, L. M., & Kc, D. B. (2021). A deep learning based approach for prediction of Chlamydomonas reinhardtii phosphorylation sites. *Scientific Reports*, *11*(1), 12550. https://doi.org/10.1038/s41598-021-91840-w

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., … Zhang, J. (2023). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Toruño, T. Y., Stergiopoulos, I., & Coaker, G. (2016). Plant-pathogen effectors: Cellular probes interfering with plant defenses in spatial and temporal manners. *Annual Review of Phytopathology*, *54*(1), 419–441. https://doi.org/10.1146/annurev-phyto-080615-100204

Vu, L. D., Gevaert, K., & De Smet, I. (2018). Protein language: Post-translational modifications talking to each other. *Trends in Plant Science*, *23*(12), 1068–1080. https://doi.org/10.1016/j.tplants.2018.09.004

Wang, Y., Li, L., Ye, T., Lu, Y., Chen, X., & Wu, Y. (2013). The inhibitory effect of ABA on floral transition is mediated by ABI5 in Arabidopsis. *Journal of Experimental Botany*, *64*(2), 675–684. https://doi.org/10.1093/jxb/ers361

Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., & Xu, D. (2017). MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, *33*(24), 3909–3916. https://doi.org/10.1093/bioinformatics/btx496

Wang, X., Zhang, Z., Zhang, C., Meng, X., Shi, X., & Qu, P. (2022). TransPhos: A deep-learning model for general phosphorylation site prediction based on transformer-encoder architecture. *International Journal of Molecular Sciences*, *23*(8), 4263. https://doi.org/10.3390/ijms23084263

Wei, L., Xing, P., Tang, J., & Zou, Q. (2017). PhosPred-RF: A novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Transactions on Nanobioscience*, *16*(4), 240–247. https://doi.org/10.1109/TNB.2017.2661756

Weissenow, K., Heinzinger, M., & Rost, B. (2022). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. *Structure*, *30*(8), 1169–1177.e4. https://doi.org/10.1016/j.str.2022.05.001

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., Harper, L. C., Gardiner, J. M., Schaeffer, M. L., & Andorf, C. M. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*, *21*(1), 385. https://doi.org/10.1186/s12870-021-03173-5

Xiong, F., Zhang, R., Meng, Z., Deng, K., Que, Y., Zhuo, F., Feng, L., Guo, S., Datla, R., & Ren, M. (2017). Brassinosteriod Insensitive 2 (BIN2) acts as a downstream effector of the target of rapamycin (TOR) signaling pathway to regulate photoautotrophic growth in Arabidopsis. *New Phytologist*, *213*(1), 233–249. https://doi.org/10.1111/nph.14118

Xue, H., Zhang, Q., Wang, P., Cao, B., Jia, C., Cheng, B., Shi, Y., Guo, W.-F., Wang, Z., Liu, Z.-X., & Cheng, H. (2022). qPTMplants: An integrative database of quantitative post-translational modifications in plants. *Nucleic Acids Research*, *50*(D1), D1491–D1499. https://doi.org/10.1093/nar/gkab945

Yang, S., Vanderbeld, B., Wan, J., & Huang, Y. (2010). Narrowing down the targets: Towards successful genetic engineering of drought-tolerant crops. *Molecular Plant*, *3*(3), 469–490. https://doi.org/10.1093/mp/ssq016

Yao, E., Blake, V. C., Cooper, L., Wight, C. P., Michel, S., Cagirici, H. B., Lazo, G. R., Birkett, C. L., Waring, D. J., Jannink, J.-L., Holmes, I., Waters, A. J., Eickholt, D. P., & Sen, T. Z. (2022). GrainGenes: A data-rich repository for small grains genetics and genomics. *Database*, *2022*, *2022*, baac034. https://doi.org/10.1093/database/baac034

Yates, A. D., Allen, J., Amode, R. M., Azov, A. G., Barba, M., Becerra, A., Bhai, J., Campbell, L. I., Carbajo Martinez, M., Chakiachvili, M., Chougule, K., Christensen, M., Contreras-Moreira, B., Cuzick, A., Da Rin Fioretto, L., Davis, P., De Silva, N. H., Diamantakis, S., Dyer, S., … Flicek, P. (2022). Ensembl genomes 2022: An expanding genome resource for non-vertebrates. *Nucleic Acids Research*, *50*(D1), D996–D1003. https://doi.org/10.1093/nar/gkab1007

Yu, K., Wang, Y., Zheng, Y., Liu, Z., Zhang, Q., Wang, S., Zhao, Q., Zhang, X., Li, X., Xu, R.-H., & Liu, Z.-X. (2023). qPTM: An updated database for PTM dynamics in human, mouse, rat and yeast. *Nucleic Acids Research*, *51*(D1), D479–D487. https://doi.org/10.1093/nar/gkac820

Zhang, W. J., Zhou, Y., Zhang, Y., Su, Y. H., & Xu, T. (2023). Protein phosphorylation: A molecular switch in plant signaling. *Cell Reports*, *42*(7), 112729. https://doi.org/10.1016/j.celrep.2023.112729

Zhao, Q., & Guo, H.-W. (2011). Paradigms and paradox in the ethylene signaling pathway and interaction network. *Molecular Plant*, *4*(4), 626–634. https://doi.org/10.1093/mp/ssr042

Zhou, Z., Yeung, W., Gravel, N., Salcedo, M., Soleymani, S., Li, S., & Kannan, N. (2023). Phosformer: An explainable transformer model for protein kinase-specific phosphorylation predictions. *Bioinformatics*, *39*(2), btad046. https://doi.org/10.1093/bioinformatics/btad046

Zuallaert, J., Ramasamy, P., Bouwmeester, R., Callewaert, N., & Degroeve, S. (2022). *PhosphoLingo: Protein language models for phosphorylation site prediction [preprint]*. Bioinformatics. https://doi.org/10.1101/2022.11.28.518163

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Poretsky, E., Andorf, C. M., & Sen, T. Z. (2023). PhosBoost: Improved phosphorylation prediction recall using gradient boosting and protein language models. *Plant Direct*, *7*(12), e554. https://doi.org/10.1002/pld3.554